CHENNAI MATHEMATICAL INSTITUTE

Reinforcement learning

Date: May6.
Due: 25th May, 2024

(1) Consider the following two policy gradient algorithms, the first called REINFORCE, which was discussed in class. The second is called baseline REINFORCE.

REINFORCE

1. Initialize θ arbitrarily.
2. For each episode do
3. Generate an episode $S_0, A_0, R_0, \ldots, S_{T-1}, A_{T-1}, R_{T-1}$, using policy parameters θ.
4. $\nabla J(\theta) = \sum_{t=0}^{t=T-1} \gamma^t G_t \frac{\partial ln(\pi(S_t, A_t, \theta))}{\partial \theta}$;
5. $\theta \leftarrow \theta + \alpha \nabla J(\theta)$;

Baseline REINFORCE

1. Initialize θ,w arbitrarily.
2. For each episode
3. Generate $S_0, A_0, R_0, \ldots, S_{T-1}, A_{T-1}, R_{T-1}$ using policy parameters θ.
4. $\nabla J(\theta) = 0$;
5. $e \leftarrow 0$;
6. for t= 0 to T-1 do
7. $\nabla J(\theta) = \nabla J(\theta) + \gamma^t (G_t - v_w(S_t)) \frac{\partial ln(\pi(S_t, A_t, \theta))}{\partial \theta}$;
8. $e \leftarrow \gamma \lambda e + \frac{\partial v_w(S_t)}{\partial w}$;
9. $\delta \leftarrow R_t + \gamma v_w(S_{t+1}) - v_w(S_t)$;
10. $w \leftarrow w + \alpha \delta e$;
11. $\theta \leftarrow \theta + \beta \nabla J(\theta)$.

When using these algorithms for updating the gradient some authors use the term γ^t and many do not. So try with both. Implement both these algorithms and apply them to the gridworld problem and the mountain car problems from assignment 1 and draw the learning curves. Also hand in a written (scanned) note of the policy representation you used and the hyperparameters used.